Name: Brownwell Ehimare

Date:29/02/2024

Course: Computer Science

**Coursework 1**

Subtask 1:

Before conducting the k-means, perform the following pre-processing tasks: scaling and outliers detection/removal and briefly justify your answer. (Suggestion: the order of scaling and outliers removal is important. The outlier removal topic is not covered in tutorials, so you need to explore it yourself). Obviously, you can implement this clustering task without exploring this “outlier ”component, however, you will not be awarded the allocated marks for this comp

1)Use box-plot for outlier Detection and Diagram Analysis:

538 outlier detected

A screen shot of a computer

Description automatically generated

After doing Experiment with scaling before outlier detection it severely affected the data by increasing the minimum value of the data higher than normal,

I used outlier detection before scaling, to adequately remove the outlier and not make them merge and severely affect other data

A screen shot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A white background with black text

Description automatically generated

**code in removing the outlier:**

**A screenshot of a computer program

Description automatically generated**

**Output:**

Number of rows removed due to outliers: 101

**Scaling the data(code):**

**A screenshot of a computer code

Description automatically generated**

**Scaling the data(output):**

As you can see from the diagram below that 101 outliers and NA values was removed from the data making the data reduced to “2599”

**A screenshot of a computer code

Description automatically generated**

Subtask 2:

Determine the number of cluster centres by showing all necessary steps/methods via “automated “tools (1.5 mark for each one of these “automated” tools)

**The four automated tool: Nbclust . Elbow Method ,Average Silhouette Method ,Gap Statistics**

**A computer screen shot of a code

Description automatically generated**

**NBCLUST(OUTPUT):2**

**A graph of a number of clusters

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**ELBOW METHOD(OUTPUT):2**

**A graph of numbers and a number of squares

Description automatically generated**

**AVERAGE SHILEOUTTE PLOT(OUTPUT): 2**

**A graph with a line and a number of clusters

Description automatically generated**

**GAT STATISTICS(OUTPUT):3**

**A graph with blue lines and numbers

Description automatically generated**

This gives a majority decision of ” k=2”

**Subtask 3:**

K-means analysis for the chosen k (all attributes used) and show all requested outputs, Show the silhouette plot (3 marks) and provide related discussion on this output, following this Kmeans attempt:

A white background with black text

Description automatically generated

A graph showing different colored shapes

Description automatically generated with medium confidence

Plot 2;

A blue and red dots

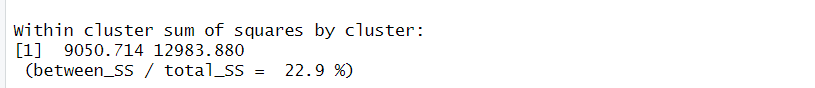
Description automatically generated

Silhouette plot(code):

A screenshot of a computer program

Description automatically generated

Silhouette plot and others(output):



A screenshot of a computer

Description automatically generated

A graph of a graph

Description automatically generated with medium confidence

A red circle with black lines

Description automatically generated

**Subtask 4:**

Apply a PCA for this wine dataset and show all related R-outputs (4 marks). Create a new dataset with those PCs with a cumulative score at least > 85%, as attributes and provide a discussion for your choice (4 marks).

A computer code with text

Description automatically generated

A computer screen shot of numbers

Description automatically generated

Given the code above I chose the PCA 1 to 7 has they gave a proportion variance of over 86% (0.3173+0.1430+0.1114+0.09727+0.0868+ 0.06714+0.06592===~88%

),There is no need to go beyond those seven has they satisfy the conditions above and they still retain relevant information’s

**Subtask 5:**

Determine the number of cluster centres by showing all necessary steps/methods via “automated”tools (1.5 mark for each one of these “automated” tools)

Code Implementation:

A screenshot of a computer program

Description automatically generated

**NBCLUST AFTER PCA:**

**A graph of different differences

Description automatically generated with medium confidence**

**A screenshot of a computer

Description automatically generated**

**ELBOW METHOD AFTER PCA:**

**A graph of numbers and a number of squares

Description automatically generated**

**AVERAGE SHILEOUTTUE AFTER PCA:**

**A graph with a line and numbers

Description automatically generated**

**GAP STATISTICS AFTER PCA:**

**A graph with numbers and lines

Description automatically generated**

This gives a total majority of kmeans=2 after pca

**SubTask 6:**

K-means analysis for this “pca”-based dataset for the chosen k and show all requested outputs 58. Show the silhouette plot (3 marks) and provide related discussion on this output, following this“pca-based” Kmeans attempt (2 marks)

A screenshot of a computer program

Description automatically generated

Correct\_k2=kmeans

K means Analysis 1

A graph showing a number of numbers

Description automatically generated with medium confidence

K means Analysis 2

A red and blue dots

Description automatically generated

Kmeans Analysis 3:

A graph of a diagram

Description automatically generated with medium confidence

**SubTask 7:**

Show the silhouette plot (3 marks) and provide related discussion on this output, following this“pca-based” Kmeans attempt

A screenshot of a computer program

Description automatically generated

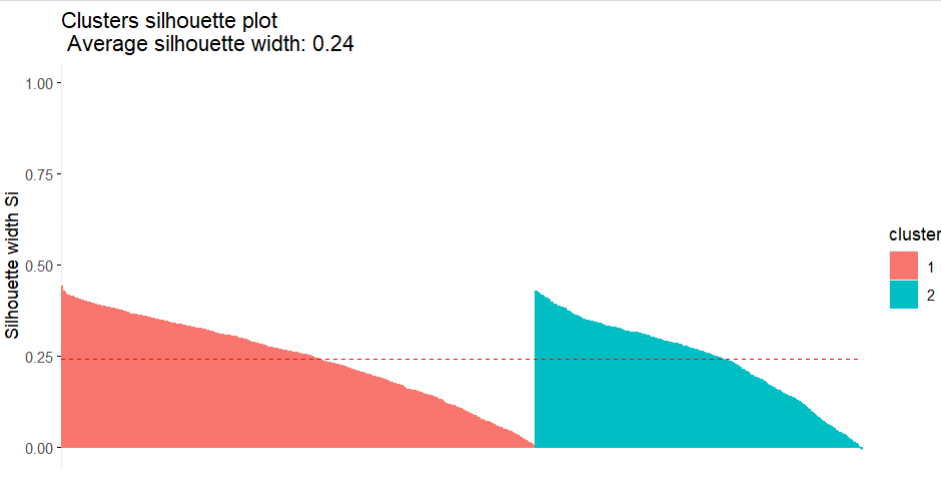
**A screenshot of a computer

Description automatically generated**

A screenshot of a computer

Description automatically generated

Silhouette Output after PCA:



**SubTask 8:**

Reference:

Wikipedia contributors. (2024,March,19). Calinski–Harabasz index. In Wikipedia. Retrieved (2024,April,26th), from <https://en.wikipedia.org/wiki/Calinski%E2%80%93Harabasz_index>

Implement and show the Calinski-Harabasz index. Provide, a brief discussion on the outcome of this index.

**////**

The calinski-harabasz index is a clustering technique for finding the optimal number of cluster

Procedures:

1. Perform clustering for different values of *k*.
2. Compute the CH index for each clustering result.
3. The value of *k* that yields the maximum CH index is chosen as the optimal number of clusters.

It is also defined as the ratio of BSS/WSS

CH=BSS/WSS

**A screenshot of a computer code

Description automatically generated**

After Implementation of the Calinski-Harabasz index

The best optimal k is 2 just like the previous techniques

Overall the best Kmeans for the data is 2 before and after PCA

A graph of a number of clusters

Description automatically generated

**CourseWork2:**

Brief discussion of the various methods used for defining the input vector in exchange rates forecasting problems 5

The method used for defining the input vector

Is a time-lagged data regression model, it operates by predicting data used 3 day prior or 2 days prior as seen below and it is very efficient

2. Evidence of various adopted input vectors and the related input/output matrices for this “AR” based approach 4

Input vector of 4

A computer code with text

Description automatically generated

A computer screen shot of text

Description automatically generated

Outputs:

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

3. Evidence of correct normalisation (3 marks) /de-normalisation (3 marks) and brief discussion of its necessity for MLP networks (3 marks) 9

T-4 normalisation

A screen shot of a computer code

Description automatically generated

T-3 normalisation

A computer code with text

Description automatically generated with medium confidence

Output:

T4 normalisation:

A screenshot of a computer screen

Description automatically generated

T3 normalisation:

A screenshot of a computer

Description automatically generated

4. Implement a number of MLPs for the “AR” approach, using various internal structures(layers/nodes)/input variables/network parameters and show in the comparison,

table, their performances(based on testing data) through the provided stat. indices. (4 marks for structures with different input vectors, 8 marks for different internal NN structures). 12,

6. Creation of the comparison matrix for this exchange rates case 4

T4:

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A close up of a text

Description automatically generated

Ouput of t4:

A close up of a screen

Description automatically generated

A screenshot of a computer screen

Description automatically generated

Tlevel3:

A white background with black text

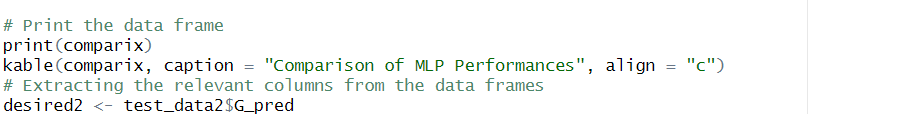
Description automatically generated

A screenshot of a computer

Description automatically generated

A white screen with black text

Description automatically generated with medium confidence



1. Discussion of the meaning of these four stat. indices (2 marks for each index) 8

**MAPE (Mean Absolute Percentage Error):** MAPE measures the average magnitude of errors as a percentage of the actual values. It provides the average percentage difference between predicted and actual values. It is calculated as the average of the absolute percentage differences between predicted and actual values.

**RMSE (Root Mean Squared Error):** RMSE measures the average magnitude of the errors between predicted and actual values. It represents the square root of the average of the squared differences between predicted and actual values. RMSE gives more weight to larger errors.

**MAE (Mean Absolute Error):** MAE measures the average magnitude of errors between predicted and actual values. It represents the average of the absolute differences between predicted and actual values. MAE is less sensitive to outliers compared to RMSE.

**SMAPE (Symmetric Mean Absolute Percentage Error):** SMAPE is similar to MAPE but considers the relative error symmetrically around zero. It calculates the average of the absolute percentage differences between predicted and actual values relative to their average. SMAPE is commonly used in forecasting tasks.

7. Discuss the issue of “efficiency” with your two best NN structures 4

T4:

The best layer is c(8,4) logistics=>

A screenshot of a computer code

Description automatically generated

2)C(2,4) logistics;  
T3)

Same for t-4 level and t3 levels

A computer screen shot of a error

Description automatically generated

8. Provide your best results both graphically (your prediction output vs. desired output)and via performance indices (2 marks for the graphical display and 2 marks for showing he requested statistical indices)

**A computer screen shot of numbers and symbols

Description automatically generated**

**A computer code with numbers and letters

Description automatically generated**